

(19)



Europäisches Patentamt
European Patent Office
Office européen des brevets



(11) Publication number:

0 492 459 A2

(12)

EUROPEAN PATENT APPLICATION(21) Application number: **91121836.0**(51) Int. Cl.⁵: **G10L 9/14**(22) Date of filing: **19.12.91**(30) Priority: **20.12.90 IT 6802990**(43) Date of publication of application:
01.07.92 Bulletin 92/27(84) Designated Contracting States:
AT BE CH DE ES FR GB GR IT LI NL SE(71) Applicant: **SIP Società Italiana per l'Esercizio
Telefonico p.a.
Via San Dalmazzo, 15
I-10122 Turin(IT)**(72) Inventor: **Drogo de Iacovo, Rosario
Via Nicola Gianniti, 7****Rocca Imperial Marina (CS)(IT)**Inventor: **Montagna, Roberto****Via Arona, 8****Torino(IT)**Inventor: **Sereno, Daniele****Via Isernia, 7/A****Torino(IT)**(74) Representative: **Riederer Freiherr von Paar zu
Schönau, Anton et al
Lederer, Keller & Riederer, Freyung 615,
Postfach 26 64
W-8300 Landshut(DE)**(54) **System for embedded coding of speech signals.**

(57) The set of possible excitation signals is subdivided into a plurality of subsets, the first of which provides the contribution to the coded signal necessary to set up a transmission at a minimum rate guaranteed by the network, whilst the others supply a contribution which, when added to that of the first subset, causes a rate increase by successive steps.

At the receiving side, a decoded signal is generated by using the excitation contribution of the first subset alone if the coded signals are received at the minimum rate, whilst for rates higher than the minimum rate the contributions of the subsets which have allowed such rate increase are also used.

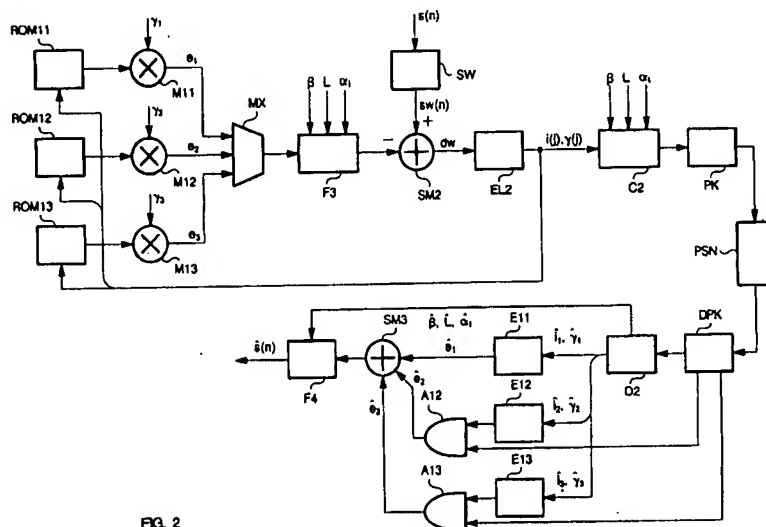


FIG. 2

The present invention concerns speech signal coding systems, and more particularly a digital coding system with embedded subcode using analysis by synthesis techniques.

The expression "digital coding with embedded subcode", or more simply "embedded coding", indicates that within a bit flow forming the coded signal, there is a slower flow which can be still decoded giving an approximate replica of the original signal. Said codes allow coping not only with accidental losses of part of the transmitted bit flow, but also with the necessity of temporary limiting the amount of information transmitted. The latter situation can occur in case of overload in packet-switched networks, e.g. those based on the so-called "Asynchronous Transfer Mode" better known as ATM, where a rate limitation can be achieved by dropping a number of packets or of bits in each packet. By using an embedded code, at the destination node the original signal is recovered, even though at the expenses of a certain degradation in comparison with the case of reception of the whole bit or packet flow. This solution is simpler than using a set of coders/decoders with different structure, operating at suitable rates and driven by network signaling for the choice of the transmission rate.

Among the systems used for speech signal coding, PCM (and more particularly uniform PCM with sample sign and magnitude coding) is per se an embedded code, since the use of a greater or smaller number of bits in a codeword determines a more or less precise reconstruction of the sample value. Other systems, such as e.g. DPCM (differential PCM) and ADPCM (adaptive differential PCM), where the past information is exploited to decode the current information, or systems based on vector quantization, such as analysis-by-synthesis coding systems, are not in their basic form embedded codings, and actually the loss of a certain number of coding bits causes a dramatic degradation in the reconstructed signal quality.

Coding-decoding devices based on DPCM or ADPCM techniques modified so as to implement an embedded coding are described in the literature. E. g., the paper entitled "Embedded DPCM for variable bit rate transmission" presented by D. J. Goodman at the Conference ICC-80, paper 42-2, describes a DPCM coder-decoder in which the signal to be coded is quantized with such a number of levels as to produce the nominal transmission rate envisaged on the line, whilst the inverse quantizers operate with the number of levels corresponding to the minimum transmission rate envisaged. The predictors in the coder and decoder operate consequently on identical signals, quantized with the same quantization step. The resulting quality degradation has proved lower than that oc-

curing in case of loss of the same number of bits in conventional DPCM coding transmission. The paper also suggests the use of the same concept for speech packet transmission, since bit dropping causes a much lower degradation than packet loss, which is the way in which usually a transmission rate is reduced under heavy traffic conditions.

In the paper entitled "Missing packet recovery of low-bit-rate coded speech using a novel packet-based embedded coder", presented by M. M. Lara-Barron and G. B. Lockhart at the Fifth European Signal Processing Conference (EUSIPCO-90), Barcelona, 18-21 September 1990, a speech signal embedded coding system is disclosed which is just studied for packet transmission in order to limit degradation in case of loss or dropping of entire packets instead of individual bits. The general coder structure basically reproduces that of the embedded DPCM coder described in the above-mentioned paper by D. J. Goodman. The system is based on a classification of packets as "essential" and "supplementary" and the network, in case of overload, preferentially drops supplementary packets. For such a classification, a current packet is compared with its prediction to determine the degradation which would result from reconstruction at the receiver, the degradation being expressed by a "reconstruction index". The reconstruction index is then compared to a threshold. If the comparison indicates high degradation, i.e. a packet difficult to reconstruct, the packet is classified as "essential", otherwise it is classified as "supplementary". The two packet types are coded and transmitted normally through the network. The decision "essential packet" or "supplementary packet" determines the position of suitable switches in the transmitter and receiver in such a manner that, at the transmitter, after transmission of a supplementary packet, the predicted packet is coded instead of the original one, and the coded packet is also supplied to a local decoder and a local predictor in order to predict the subsequent packet. At the receiver, essential packets are decoded normally and supplied to the output. A local encoder is also provided for updating the decoder parameters in case of a missing packet, by using a packet predicted in a local predictor. A supplementary packet is decoded and emitted normally, but it is supplied also to the local predictor and encoder to keep the encoder parameters in alignment with the encoder parameters at the transmitter.

DPCM/ADPCM coding systems offer good performance for rates basically comprised in the interval 32 to 64 kbit/s, while at lower rates their performance strongly decreases as the rate decreases. At lower rates different coding techniques are used, more particularly analysis-by-synthesis techniques. Yet, also these techniques do not result in embed-

ded codes, neither does the literature describe how an embedded code can be obtained. The paper by M. M. Lara-Barron and G. B. Lockhart states that the suggested method can also be applied to any low-bit rate encoder that utilises past information to decode current-frame samples, and hence theoretically such a method could be used also in case of analysis-by-synthesis coding techniques. However, even neglecting the fact that indications of performance are given only for 32 kbit/s ADPCM coding, the structure of transmitter and receiver is the typical structure of DPCM/ADPCM systems, comprising, in addition to the actual coding circuits at the transmitter and decoding circuits at the receiver, a decoder and a predictor at the transmitter and a predictor at the receiver: said devices are not provided for in the transmitters/receivers of a system exploiting analysis-by-synthesis techniques, and their addition, besides that of the circuits for determining the reconstruction-index, would greatly complicate the structure of said transmitters/receivers. Furthermore, since the coding/decoding circuits comprise a certain number of digital filters, the problem arises of correctly updating their memories.

The present invention provides a method of and a device for speech signal coding, allowing attainment of an embedded coding when using analysis-by-synthesis techniques, while keeping the typical structure of the transmitters/receivers of such systems unchanged.

The method comprises a coding phase, in which at each frame a coded signal is generated which comprises information relevant to an excitation, chosen out of a set of possible excitation signals and submitted to a synthesis filtering to introduce into the excitation short-term and long-term spectral characteristics of the speech signal and to produce a synthesized signal, the excitation chosen being that which minimises a perceptually-significant distortion measure, obtained by comparison of the original and synthesized signals and simultaneous spectral shaping of the compared signals, and a decoding phase wherein an excitation, chosen according to the information contained in a received coded signal out of a signal set identical to the one used for coding, is submitted to a synthesis filtering corresponding to that effected on the excitation during the coding phase, and is characterised in that, to implement an embedded coding for use in a network where the coded signals are organised into packets which are transmitted at a first bit rate and can be received at bit rates lower than the first rate but not lower than a predetermined minimum transmission rate, the various rates differing by discrete steps:

- the sets of excitation signals for coding and decoding are split into a plurality of subsets,

the first of which contributes to the respective excitation with such an amount of information as required for a transmission of the coded signals at the minimum transmission rate, whilst the other subsets provide contributions corresponding each to one of said discrete steps, the contributions of said other subsets being used in a predetermined succession and being added to the contributions of the first subset and of previous subsets in the succession;

- during the coding phase the contributions supplied by all subsets of excitation signals are filtered in such a manner that, at each frame, the memory of the filtering results relevant to one or more preceding frames is taken into account only when filtering the excitation contribution of the first subset, whilst the excitation contributions of all other subsets are filtered without taking into account the results of the filtering relevant to preceding frames;
- still during the coding phase, the contributions to the coded signal supplied by different subsets are inserted into different packets which can be distinguished from one another, the decrease from the first rate to one of the lower rates being achieved by first discarding packets containing the excitation contribution which has led to the attainment of the first rate and then packets containing the excitation contribution corresponding to preceding increase steps;
- during the decoding phase, for each frame, the excitation contributions of the first subset are submitted to the synthesis filtering whatever the bit rate at which the coded signals are received and, if such a rate is higher than the minimum rate, even excitation contributions of the subsets corresponding to the steps which have led to such a rate, are filtered, the filtering of the excitation signals in the first subset being a filtering with memory and the filtering of the excitation signals in the other subsets being a filtering without memory. A device for implementing the method comprises a coder including:
 - a first excitation source supplying a set of excitation signals wherein an excitation to be used for coding operations relevant to a frame of samples of the speech signal is chosen;
 - a first filtering system which imposes on the excitation signals the short-term and long-term spectral characteristics of the speech signal and supplies a synthesized signal;
 - means for carrying out a perceptually significant measurement of the distortion of the

synthesized signal in comparison with the speech signal, for searching an optimum excitation which is the excitation which minimises the distortion, and for generating coded signals comprising information relevant to the optimum excitation signal; and

- means to organise a transmission of coded signals as a packet flow;

and a decoder including:

- means for extracting the coded signals from a received packet flow;
- a second excitation source supplying a set of excitation signals corresponding to the set supplied by the first source, an excitation corresponding to the one used for coding during a frame being chosen in said set on the basis of the excitation information contained in the coded signal; and
- a second filtering system, identical to the first one, which generates a synthesized signal during decoding;

and is characterised in that:

- the first source of excitation signals comprises a plurality of partial sources each arranged to supply a different subset of the excitation signals, the subset supplied by a first partial source contributing to the coded signal with a bit stream necessary to obtain a packet transmission at a minimum bit rate, while the subsets of the other partial sources contribute to the coded signal with bit streams that, successively added to the contribution supplied by the first partial source, originate an increase of the bit rate by discrete steps up to a maximum bit rate;
- the second source of excitation signals comprises a plurality of partial sources supplying respective subsets of the excitation signals corresponding to the subsets supplied by the partial sources of the first excitation signals;
- the first and second filtering systems comprise each a first filtering structure which is fed with the excitation signals belonging to the first subset and, during the filtering relevant to a frame, processes them exploiting the memory of the filterings relevant to preceding frames, and further filtering structures, which are each associated with one of the other subsets of excitation signals and which, during the filterings relevant to a frame, process the relevant signals without exploiting the memory of the filtering relevant to the preceding frames;
- the means for measuring distortion and searching the optimum excitation supply the means generating the coded signal with an excitation comprising contributions from all subsets of excitation signals;

- the means for organising the transmission into packets introduce into different packets the excitation information originating from different subsets of excitation signals; and
- the second filtering system supplies the signal synthesized during decoding by processing an excitation always comprising a contribution from the first subset of excitation signals, and comprising contributions from one or more further subsets only if the packet flow relevant to a frame of samples of speech signal is received at higher rate than the minimum rate.

Coding systems using CELP (Codebook Excited Linear Prediction) technique, which is an analysis-by-synthesis technique, are also known, where the excitation codebook is subdivided into partial codebooks. An example is described by I. A. Gerson and M. A. Jasuk in the paper entitled: "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 kbps" presented at the International Conference on Acoustics, Speech and Signal Processing (ICASSP 90), Albuquerque (USA), 3 - 6 April 1990. However, these systems are employed in fixed rate networks, and hence also at the receiving side the excitation always comprises contributions of all partial codebooks and the problem of tuning the filters at the transmitter and at the receiver does not exist.

The invention also provides a method of transmitting signals coded by analysis-by-synthesis techniques with the coding method and the coding device according to the invention. The invention will become more apparent with reference to the annexed drawings, which show the implementation of the invention in case of use of CELP technique and in which:

- Fig. 1 is a basic diagram of a conventional CELP coder;
- Fig. 2 is a basic diagram of a coder according to the invention;
- Fig. 3 and Fig. 4 are basic diagrams of the filtering system of the receiver and transmitter of the system of Fig. 2;
- Fig. 5 is a functional diagram of the filtering system in the transmitter;
- Fig. 6 is a partial diagram of a variant.

Prior to describing the invention, we will shortly disclose the structure of a speech-signal CELP coding/decoding system. As known, in such systems the excitation signal for the synthesis filter simulating the vocal tract consists of vectors, obtained e.g. from random sequences of Gaussian white noise, chosen out of a convenient codebook. During the coding phase, for a given block of speech signal samples, the vector is to be looked for which, supplied to the synthesis filter, minimises a perceptually-significant distortion mea-

sure, obtained by comparing the synthesized samples and the corresponding samples of the original signal, and simultaneous weighting by a function which takes into account also how human perception evaluates the distortion introduced. This operation is typical of all systems based on analysis-by-synthesis techniques, which differ in the nature of the excitation signal.

With reference to Fig. 1, the transmitter of a CELP coding system can be schematized by:

- a filtering system F1 (synthesis filter) simulating the vocal tract and comprising the cascade of long-term synthesis filter (predictor) LT1 and of a short-term synthesis filter (predictor) ST1, which introduce into the excitation signal the characteristics depending on the fine spectral structure of the signal (more particularly the periodicity of voiced sounds) and those depending on the spectral envelope of the signal, respectively. A typical transfer functions for the long term filter is

$$B(z) = 1/(1 - \beta z^{-L}) \quad (1)$$

where z^{-1} is a delay by one sampling interval, β and L are the gain and the delay of the long-term synthesis (the latter being the pitch period or a multiple thereof in case of voiced sounds). A typical transfer function for the short-term filter is

$$A(z) = 1/(1 - \sum \alpha_i z^{-i}) \quad (2)$$

where α_i is a vector of linear prediction coefficients, determined from input signal $s(n)$ using the well known linear prediction techniques, and the summation extends to all samples in the block;

- a read only memory ROM1 which contains the codebook of vectors (or words), which, weighted by a scale factor γ in a multiplier M, form the excitation signal $e(n)$ to be filtered in F1; a same scale factor, previously determined, can be used for the whole search for an optimum vector (i. e. the vector minimizing the distortion for the block of samples being coded), or an optimum scale factor for each vector can be determined and used during the search;
- an adder SM1, which carries out the comparison between the original signal $s(n)$ and the filtered signal $s1(n)$ and supplies an error signal $d(n)$ consisting of the difference between said two signals;
- a filter SW1 for spectrally shaping the error signal, so as to render the differences between the original and the reconstructed signal less perceptible; typically SW1 has a

transfer function of the type

$$W(z) = (1 - \sum \alpha_i z^{-i}) / (1 - \sum \alpha_i \lambda^i z^{-i}) \quad (3)$$

where λ is an experimentally determined constant corrective factor (typically, of the order of 0.8 - 0.9) which determines the band increase around the formants; this filter could be located upstream SM1, on both inputs, so that SM1 directly gives the weighted error: in such case, the transfer function of ST1 becomes $1 / (1 - \sum \alpha_i \lambda^i z^{-i})$;

- a processing unit EL1 which carries out the operation necessary for searching the optimum excitation vector and possibly optimizing the scale factor and the long-term filter parameters.

The coded signal, for each block, consists of index i of the optimum vector chosen, scale factor γ , delay L and gain β of LT1, and coefficients α_i of ST1, duly quantized in a coder C1. Clearly, the filters in F1 ought to be reset at each new block to be coded.

The receiver comprises a decoder D1, a second read-only memory ROM2, a multiplier M2, and a synthesis filter F2 comprising the cascade of a long-term synthesis filter LT2 and a short-term synthesis filter ST2, identical respectively to devices ROM1, M1, F1, LT1, ST1 in the transmitter. Memory ROM2, addressed by decoded index \hat{i} , supplies F2 with the same vector as used at the transmitting side, and this vector is weighted in M2 and filtered in F2 by using scale factor $\hat{\gamma}$ and parameters $\hat{\alpha}$, $\hat{\beta}$, \hat{L} , of short term and long term synthesis corresponding to those used in the transmitter and reconstructed starting from the coded signal; output signal $\hat{s}(n)$ of filter F2, converted again if necessary into analog form, is supplied to utilising devices.

In the particular case of use in an ATM network (or in general in a packet switched network) downstream the encoder there are devices for organising the information into packets to be transmitted, and upstream the decoder there are devices for extracting from packets received the information to be decoded. These devices are well known to the skilled in the art, and their operation do no affect coding/decoding operations.

Fig. 2 shows the embedded coder of the invention. By way of a non-limiting example, it will be supposed that such a coder is used in a packet switched network PSN (more particularly, an ATM network) where it is possible to drop a number of packets (independently of their nature) to reduce the transmission rate in case of overload. For simplicity and clarity of description, reference will be made to a speech coder capable of operating at 9.6, 8 or 6.4 kbit/s according to traffic conditions.

Said rates lie within the range for which analysis-by-synthesis coders are typically used.

To implement the embedded coding, the excitation codebook is split into three partial codebooks. The first partial codebook contains such a number of vectors as to contribute to the coded signal with a bit stream that, added to the bit stream produced by the coding of the other parameters (scale factor and filtering system parameters), gives rise to the minimum transmission rate of 6.4 kbit/s; the second and third partial codebooks have such a size as to provide the contribution required by a transmission rate of 1.6 kbit/s. ROM11, ROM12, ROM13 denote the memories containing the partial codebooks; M11, M12, M13 denote the multipliers that weight the codevectors by the respective scale factors γ_1 , γ_2 , γ_3 , giving excitation signals e_1 , e_2 , e_3 . The transmitter always operates at 9.6 kbit/s, and hence the coded signal comprises, as far as the excitation is concerned, the contributions provided by the three above-mentioned signals. Advantageously, to keep the total number of bits to be transmitted limited, the filtering system will be identical (i.e. it will use the same weighting coefficients) for all excitations. Therefore the Figure shows a single filter F3 connected to the outputs of multipliers M11, M12, M13 through a multiplexer MX. For drawing simplicity the two predictors in F3 have not been indicated. In the diagram it has also been supposed that spectral weighting is effected separately on input signal $s(n)$ and on the excitation signals, so that adder SM2 (analogous to SM1, Fig. 1) directly gives weighted error dw . Filter SW is hence indicated only on the path of $s(n)$, since its effect on the excitation is obtained by a suitable choice of short term synthesis filter F3, as already explained. EL2 denotes the processing unit which performs the search for the optimum vector within the partial codebooks and the operations required for optimizing the other parameters (in particular, scale factor and gain of long-term filter) according to any of the procedures known in the art. C2 denotes a device having the same functions as C1 in Fig. 1. Clearly, the coded signals will comprise indices $i(j)$ ($j = 1, 2, 3$) of the optimum vectors chosen in the three partial codebooks and the respective optimum scale factor $\gamma(j)$.

Quantizer C2 is followed by device PK packetising the coded speech signal in the manner required by the particular packet switching network PSN. The excitation contribution of the different codebooks will be introduced by PK into different packets labeled so that they can be distinguished in the different networks nodes. This can be easily obtained by exploiting a suitable field in the packet header. Thus, in case of overload, a node can drop first the packets containing the excitation contribu-

tion from e_3 and then the packets containing contribution from e_2 ; the packets with the contribution from e_1 are on the contrary always forwarded through the network, and form the minimum 6.4 kbit/s data flow guaranteed.

At the receiver, a device DPK extracts from the packets received the coded speech signals and sends them to decoding circuit D2, analogous to D1 (Fig. 1), which is connected to three sources of reconstructed excitation E11, E12, E13. Each source comprises a read-only-memory, addressed by a respective decoded index \hat{i}_1 , \hat{i}_2 , \hat{i}_3 and containing the same codebook as ROM11, ROM12 or ROM13, respectively, and a multiplier, analogous to multiplier M2 (Fig. 1) and fed with a respective decoded scale factor $\hat{\gamma}_1$, $\hat{\gamma}_2$ or $\hat{\gamma}_3$. Depending on the rate at which the speech signal is received, synthesis filter F4, analogous to filter F2 of Fig. 1, will receive the only excitation supplied by E11 (in case 6.4 Kbit/s are received) or the excitation from E11 and E12 (8 kbit/s) or the excitations supplied by E11, E12, E13 (9.6 kbit/s). This is schematized by adder SM3, which directly receives the signals from E11 and receives the output signals of E12, E13 through AND gates A12, A13 enabled e.g. by DPK when necessary.

For drawing simplicity neither the various timing signals for the transmitter and receiver components, nor the devices generating them are indicated; on the other hand timing aspects are not affected by the invention.

To keep a good quality of the reconstructed signal, the filter operation at the transmitter and the receiver must be as uniform as possible. In accordance with the invention, taking into account that at least the data flow at minimum speed is guaranteed by the network, the coder has been optimised for such minimum speed. This corresponds to carrying out coding/decoding in a frame by exploiting the memory contribution of filters F3, F4 relevant to the only first excitation, whilst the second and the third excitations are submitted to a filtering without memory. In other terms, the optimization procedure is carried out by taking into account the filterings carried out in the preceding frames for the search of a vector in ROM11, and by taking into account the only current frame for the search in ROM12, ROM13. As a consequence, even at the receiver, only the filtering of excitation signals \hat{e}_1 will take into account the results of the previous filterings.

The basic diagrams of the receiver and the transmitter under these conditions are represented in Figs. 3 and 4. For a better understanding of those diagrams and of the following ones it is to be taken into account that a digital filter with memory can be schematized by the parallel connection of two filters having the same transfer function as the

one considered: the first filter is a zero input filter, and hence its output represents the contribution of the memory of the preceding filterings, whilst the second filter actually processes the signal to be filtered, but it is initialised at each frame by resetting its memory (supposing for simplicity that the vector length coincides with the frame length). Furthermore, a filtering without memory is a linear operation, and hence the superposition of effects applies: in other terms, with reference to Fig. 2, in case of reception at a rate exceeding the minimum, filtering without memory the signal resulting from the sum of \hat{e}_1 , \hat{e}_2 , and possibly \hat{e}_3 corresponds to summing the same signals filtered separately without memory.

In Fig. 3 filtering system F4 of Fig. 2 is represented as subdivided into three subsystems F41, F42, F43 for processing excitations \hat{e}_1 , \hat{e}_2 , \hat{e}_3 , respectively. Subsystem F41 carries out a filtering with memory, and hence it has been represented as comprising zero-input element F41a and element F41b filtering excitation \hat{e}_1 without memory. The outputs of elements F41a, F41b are combined in adder SM31, whose output u1 conveys the reconstructed digital speech signal in case of 6.4 kbit/s transmission. Subsystems F42, F43 filter \hat{e}_2 , \hat{e}_3 without memory and hence are analogous to F41b. The output signal of filter F42 is combined with the signal on u1 in an adder SM32, whose output u2 conveys the reconstructed digital speech signal in case 8 kbit/s are received. Finally, the output signal of filter F43 is combined with the signal present on u2 in an adder SM33, whose output u3 conveys the reconstructed digital speech signal in case of 9.6 kbit/s transmission.

The diagram of Fig. 4 is quite similar: F31 (F31a, F31b), F32, F33 are the subsystems forming F3, and SM21, SM22, SM23, SM24 is a chain of adders generating signal dw of Fig. 2. More particularly, the output signal of F31a, i.e. the contribution of the memories of filtering of excitation e_1 , is subtracted from weighted input signal sw(n) in SM21, yielding a first partial error dw1; the output signal of F31b, i.e. the result of the filtering without memory of e_1 , is subtracted from dw1 in SM22 yielding a second partial error signal dw2; the contribution due to filtering without memory of e_2 is subtracted from dw2 in SM23, yielding a signal dw3, from which the contribution due to the filtering without memory of e_3 is subtracted in SM24. For a better understanding of the following diagrams, the cascade of long-term and short-term predictors LT31a, ST31a and LT31b, ST31b is explicitly indicated in F31a, F31b. All predictors in the various elements have transfer functions given by (1) or (2), as the case may be.

Fig. 5 shows the structure of filtering system F3, under the hypothesis that the length of a frame

coincides with the length of the vectors in the excitation codebook and that delay L of long-term predictors is greater than the vector length: this choice for the delay is usual in CELP coders. Corresponding devices are denoted by the same references in Figs. 4 and 5.

Element F31a simply comprises two short-term filters ST311, ST312 and multiplier M3, in series with ST312, which carries out the multiplication by factor β which appears in (1). Filter ST311 is a zero input filter, whilst ST312 is fed, for processing the n-th sample of a frame, with output signal PIT(n-L), relevant to L preceding sampling instants, of a long-term synthesis filter LT3' which receives the samples of e_1 (Fig. 2) and, with a short-term synthesis filter ST3', forms a fictitious synthesizer SIN3 serving to create the memories for element F31a.

This structure has the same functions as the cascade of LT31a and ST31a in Fig. 4. In fact, at instant n, a filter such as LT31a (with zero input) would supply ST31a with the filtered signal relevant to instant n-L, weighted by factor β . This same signal can be obtained by delaying the output signal of LT3' by L sampling instants in a delay element DL1, so that LT31a can be eliminated. ST31a, as disclosed above, can be split into two filters ST311, ST312 with zero input and memory and with input PIT(n-L) and without memory, respectively. The memory for ST311 will consist of output signal ZER(n) of ST3'. The output signal of ST311 is fed to the input of an adder SM211, where it is subtracted from signal sw(n), and the output signal of the cascade of ST312 and M3 is connected to an adder SM212, where it is subtracted from the output signal of SM211; the two adders carry out the functions of adder SM21 in Fig. 5.

Element F31b without memory comprises only short-term synthesis filter ST31b: in fact, with the hypothesis made for delay L, long-term synthesis filter LT31b would let through the input signal unchanged, since the output sample to be used for processing an input sample would be relevant to the preceding frames. For the same reasons, filters F32, F33 of Fig. 4 only comprise short-term synthesis filters, here denoted by ST32, ST33.

As stated, the scheme of Fig. 5 is based on the assumption that the frame length coincide with the length of the codebook vectors. Usually however the frames have a duration of the order of 20 ms (160 samples of speech signal at a sampling frequency of 8 kHz), and the use of vectors of such a length would require very big memories and give rise to high computing complexity for minimising the error. Generally it is preferred to use shorter vectors (e.g. vectors with length 1/4 of the frame duration) and subdivide the frames into subframes

of the same length as a codebook vector, so that an excitation vector per each subframe is used for the coding. Thus, during a frame, the search for the optimum vector in each partial codebook is repeated as many times as the subframes are. In an ATM network, packet dropping for limiting the transmission rate takes place when passing from one frame to the next, whilst within the frame the rate is constant. Within a frame it is then possible to optimise the coder for the rate actually used in that frame, i.e. to take also into account the memories of filters F32, F33. The long-term prediction delay will still be greater than vector duration. Under these conditions also filters F32, F33 would have the structure shown for F31 in Fig. 5, with the only difference that at the end of each frame signals PIT and ZER relevant to e_2 , e_3 will have to be reset, since only the memory of F31 is taken into account.

The structure can be simplified if long-term characteristics are not taken into account for filtering excitations e_2 , e_3 (and hence \hat{e}_2 , \hat{e}_3): in this case in fact the fictitious synthesizer relevant to each one of said excitations comprises only a short-term synthesis filter and the branch which receives signal PIT is missing. As shown in Fig. 6, under these conditions filtering subsystems F32, F33 comprise the three filters ST32a, ST32b, ST32' and ST33a, ST33b, ST33' respectively, analogous to ST311, ST31b and ST3' (Fig. 5), and adders SM231, SM232 and SM241, SM242 forming adders S23 and S24, respectively. ZER2, ZER3 denote signs corresponding to ZER (Fig. 5), i.e. signals representing the memory contribution for filtering in F32, F33; finally, RSM denotes the reset signal for the memories of ST32', ST33', which is generated at the beginning of each new frame by the conventional devices timing the operations of the coding system.

It is clear that the above description has been given only by way of a non limiting example, variations and modifications being possible without going out of the scope of the invention. More particularly, even though reference has been made to a CELP coding scheme, the invention can apply to whatever analysis-by-synthesis coding system, since the invention is per se independent of excitation signal nature. More particularly, in case of multipulse coding, which with CELP coding is the most widely used, a first number of pulses will be used to obtain 6.4 kbit/s transmission rate, and two other pulse sets will provide the rate increase required to achieve the other envisaged speeds.

Claims

1. A method of coding by analysis-by-synthesis techniques speech signals converted into

frames of digital samples, comprising a coding phase, in which at each frame a coded signal is generated comprising information relevant to an excitation, chosen out of a set of possible excitation signals and submitted to a synthesis filtering to introduce into the excitation short-term and long-term spectral characteristics of the speech signal and to produce a synthesized signal, the excitation chosen being that which minimises a perceptually-significant distortion measure obtained by comparison of the original and synthesized signals and simultaneous spectral shaping of the compared signals, and a decoding phase wherein an excitation, chosen out of a signal set identical to the one used for coding by exploiting the excitation information contained in a received coded signal, is submitted to a synthesis filtering corresponding to that effected on the excitation during the coding phase, characterised in that, to implement an embedded coding for use in a network where the coded signals are organised into packets which are transmitted at a first bit rate and can be received at bit rates lower than the first rate but not lower than a predetermined minimum transmission rate, the various rates differing by discrete steps:

- the sets of excitation signals for coding and decoding are split into a plurality of subsets, the first of which contributes to the respective excitation with such an amount of information as required for transmission of the coded signals at the minimum transmission rate, whilst the other subsets provide contributions corresponding each to one of said discrete steps, the contributions of said other subsets being used in a predetermined succession and being added to the contributions of the first subset and of preceding subsets in the succession;
- during the coding phase the contributions supplied by all subsets of excitation signals are filtered in such a manner that, at each frame, the memory of the filtering results relevant to one or more preceding frames is taken into account only when filtering the excitation contribution of the first subset, whilst the excitation contributions of all other subsets are filtered without taking into account the results of the filtering relevant to preceding frames;
- still during the coding phase, the contributions supplied by different subsets are inserted into different packets which can be distinguished from one another, the decrease from the first rate to one of the lower rates being achieved by dis-

carding first packets containing the excitation contribution which has led to the attainment of the first rate and then packets containing the excitation contribution corresponding to preceding increase steps;

- during the decoding phase, for each frame, the excitation contribution of the first subset is submitted to synthesis filtering whatever the bit rate at which the coded signal is received, and, if such a rate is higher than the minimum rate, there are filtered also excitation contributions of the subsets corresponding to the steps which have led to such a rate, the filtering of the excitation contribution of the first subset being a filtering with memory and the filtering of the excitation contributions of the other subsets being a filtering without memory.

2. A method as claimed in claim 1, wherein the excitation to be used for coding in a frame comprises a plurality of excitation signals of each subset, characterised in that during coding and decoding the filtering of an excitation signal takes into account, for all subsets, the memory of the preceding filterings of signals relevant to the same frame.

3. A method as claimed in claim 1 or 2, characterised in that the synthesis filtering introduces into the excitation the long-term characteristics only for the contribution of the first subset.

4. A device for coding and decoding speech signals by analysis-by-synthesis techniques, for implementing the method as claimed in any one of claims 1-3, comprising a coder including:

- a first excitation source (ROM11, M11, ROM12, M12, ROM13, M13) supplying a set of excitation signals (e_1 , e_2 , e_3) wherein an excitation to be used for coding operations relevant to a frame of samples of the speech signal is chosen;
- a first filtering system (F3) which imposes on the excitation signals the short-term and long-term spectral characteristics of the speech signal and supplies a synthesized signal;
- means (SW, SM2, EL2, C2) for carrying out a perceptually significant measurement of the distortion of the synthesized signal in comparison with the speech signal, for searching an optimum excitation which is the excitation minimising the distortion, and for generating coded sig-

nals comprising information relevant to the optimum excitation; and

- means (PK) to organise a transmission of coded signals as a packet flow;

and also comprising a decoder including:

- means (DPK) for extracting the coded signals from a received packet flow;
- a second excitation source (E11, E12, E13) supplying a set of excitation signals (\hat{e}_1 , \hat{e}_2 , \hat{e}_3) corresponding to the set supplied by the first source (ROM11, M11, ROM12, M12, ROM13, M13), an excitation corresponding to the one used for coding during a frame being chosen in said set on the basis of the excitation information contained in the coded signal; and
- a second filtering system (F4), identical to the first (F3), which generates a synthesized signal during decoding;

characterised in that:

- the first source of excitation signals (ROM11, M11, ROM12, M12, ROM13, M13) comprises a plurality of partial sources each arranged to supply a different subset of the excitation signals, the subset (e_1) supplied by a first partial source (ROM11, M11) contributing to the coded signal with a bit stream necessary to obtain a packet transmission at a minimum bit rate, while the subsets (e_2 , e_3) of the other partial sources (ROM12, M12, ROM13, M13) contribute to the coded signal with bit streams that, successively added to the contribution supplied by the first partial source (ROM11, M11), originate an increase of the bit rate by discrete steps up to a maximum bit rate;
- the second source of excitation signals (E11, E12, E13) comprises a plurality of partial sources supplying respective subsets of the excitation signals corresponding to the subsets supplied by the partial sources of the first excitation source;
- the first and second filtering systems (F3, F4) comprise each a first filtering structure (F31, F41) which is fed with the excitation signals belonging to the first subset (e_1 , \hat{e}_1) and, during the filtering relevant to a frame, processes them exploiting the memory of the filterings relevant to preceding frames, and further filtering structures (F32, F33; F42, F43), which are each associated with one of the other subsets of excitation signals and which, during the filterings relevant to a frame, process the relevant signals

- without exploiting the memory of the filtering relevant to the preceding frames;
- the means (SW, SM2, EL2) for measuring distortion and searching the optimum excitation supply the means (C2) generating the coded signal with an excitation comprising contributions from all subsets of excitation signals;
 - the means (PK) for organising the transmission into packets introduce into different packets the excitation information originating from different subsets of excitation signals; and
 - the second filtering system (F4) supplies the signal synthesized during decoding by processing an excitation always comprising a contribution from the first subset of excitation signals (\hat{e}_1), and comprising contributions from one or more further subsets (\hat{e}_2 , \hat{e}_3) only if the packet flow relevant to a frame of samples of speech signal is received at higher rate than the minimum rate.
5. A device as claimed in claim 4, characterised in that each subset of excitation signals contributes to the coded signal relevant to a frame with a plurality of excitation signals, and said further filtering structures (F32, F33; F42, F43) comprise memory elements for storing the results of filterings carried out on blocks of preceding samples relevant to the same frame, such memory elements being reset at the beginning of the filtering operations relevant to a new frame.
6. A device as claimed in claims 4 or 5, characterised in that the first filtering structure (F31, F41) in the coder and the decoder contains the cascade of short-term synthesis filter and a long-term synthesis filter, and the further filtering structures (F32, F33; F42, F43) consist of a short-term synthesis filter.
7. A method of transmitting packetized coded speech signals in a network where packets are transmitted at a first bit rate and can be received at a bit rate lower than the first one but not lower than a guaranteed minimum speed, the speech signals being coded with analysis by synthesis techniques in which an excitation, chosen within a set of possible excitation signals, is processed in a filtering system (F3, F4) which inserts into the excitation the long-term and short-term characteristics of the speech signal, characterised in that:
- the excitation chosen for coding at the transmitting side comprises contributions

provided by a plurality of excitation branches (ROM11, M11, ROM12, M12, ROM13, M13), the first of which (ROM11, M11) provides a contribution allowing a transmission at the minimum rate, whilst each other branch (ROM12, M12, ROM13, M13), provides the contribution necessary to increase the transmission rate, by a succession of predetermined steps, from the minimum rate to the first rate;

- during coding operations relevant to a frame of digital samples of speech signal, the excitation supplied by the first branch (ROM11, M11) is filtered taking into account the results of filterings carried out during the coding operations relevant to preceding frames and the excitation supplied by the other branches (ROM12, M12, ROM13, M13) is filtered without taking into account such results;
- the contributions supplied by different branches are inserted into different packets, labeled so as to be distinguished from one another;

in that along the network the possible packet suppression is carried out only on packets containing the excitation contributions supplied by branches different from the first one and takes place starting with those containing excitation contributions corresponding to the step which has brought the transmission rate to the first value and going on then with the packets containing excitation contribution corresponding to a preceding increase step; and in that

- the excitation to be submitted to filtering for decoding at the receiving side always comprises the contribution supplied by a first branch, corresponding to the first excitation branch at the transmitting side, and, if the bit rate at which the packets in a frame are received is higher than the minimum rate, the excitation also comprises contributions of excitation branches corresponding to increase step or steps which bring to such a rate;
- the filtering of the contributions of the different excitation branches, during decoding of the signals relevant to a frame of digital samples of speech signal to be decoded, is carried out by taking into account the results of the filtering of the signals relevant to preceding frames for the first excitation branch and without taking into account such results for the other excitation branches.

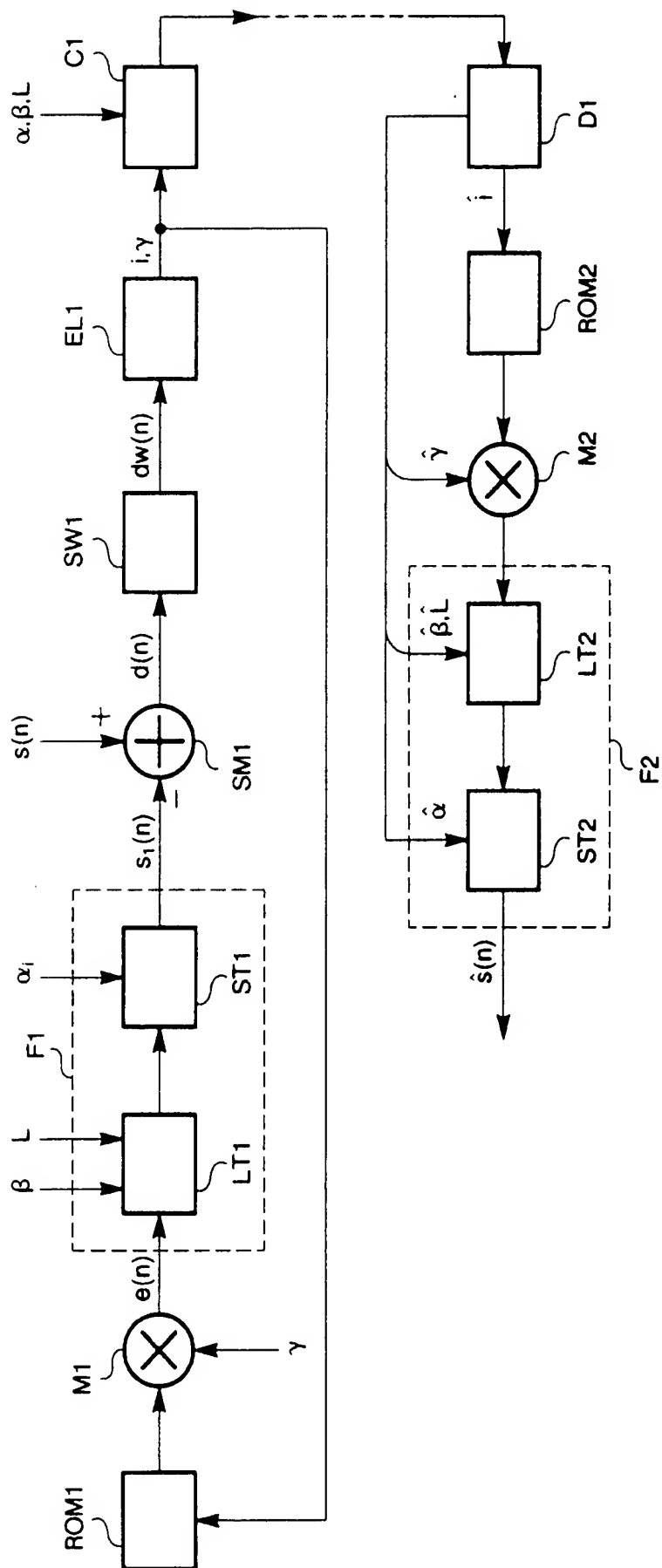


FIG. 1

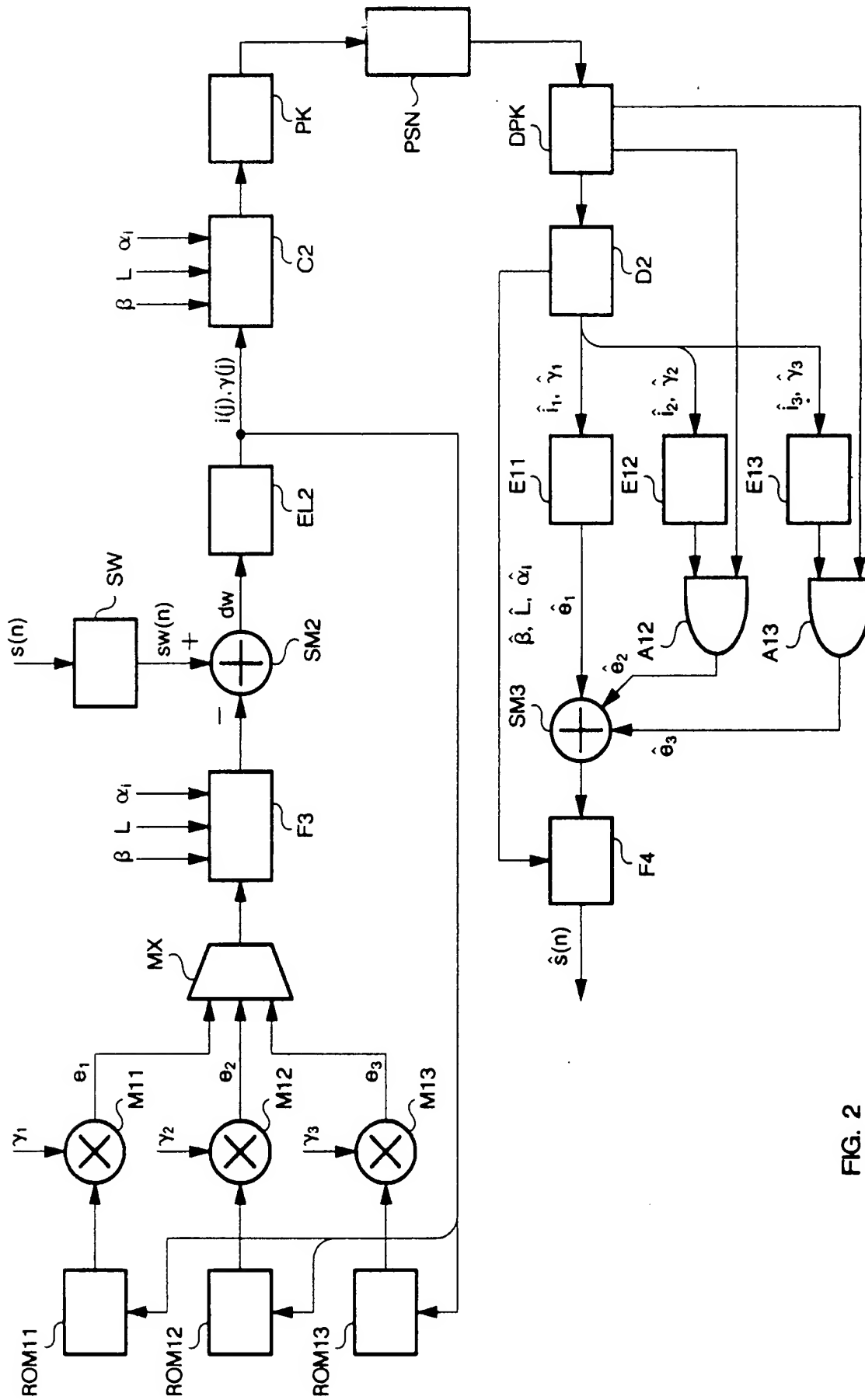


FIG. 2

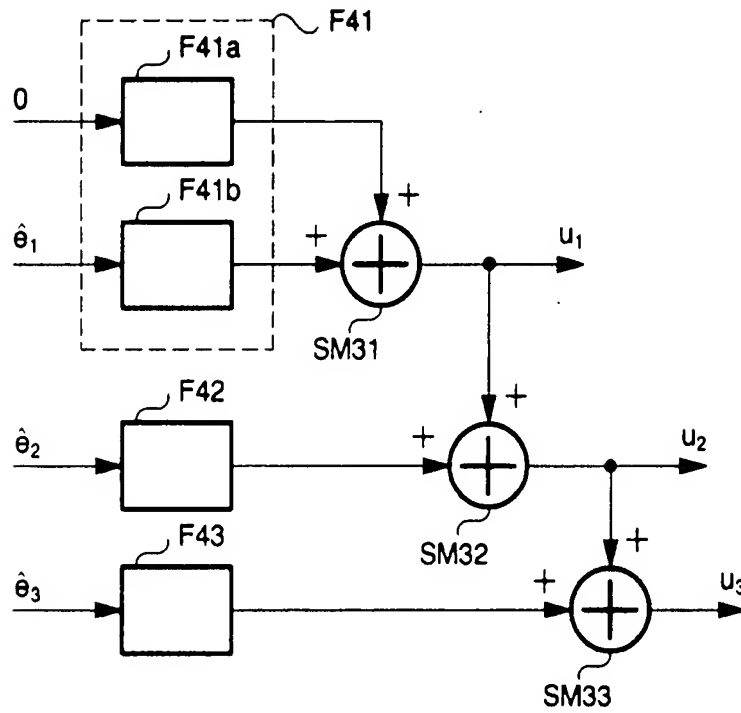


FIG. 3

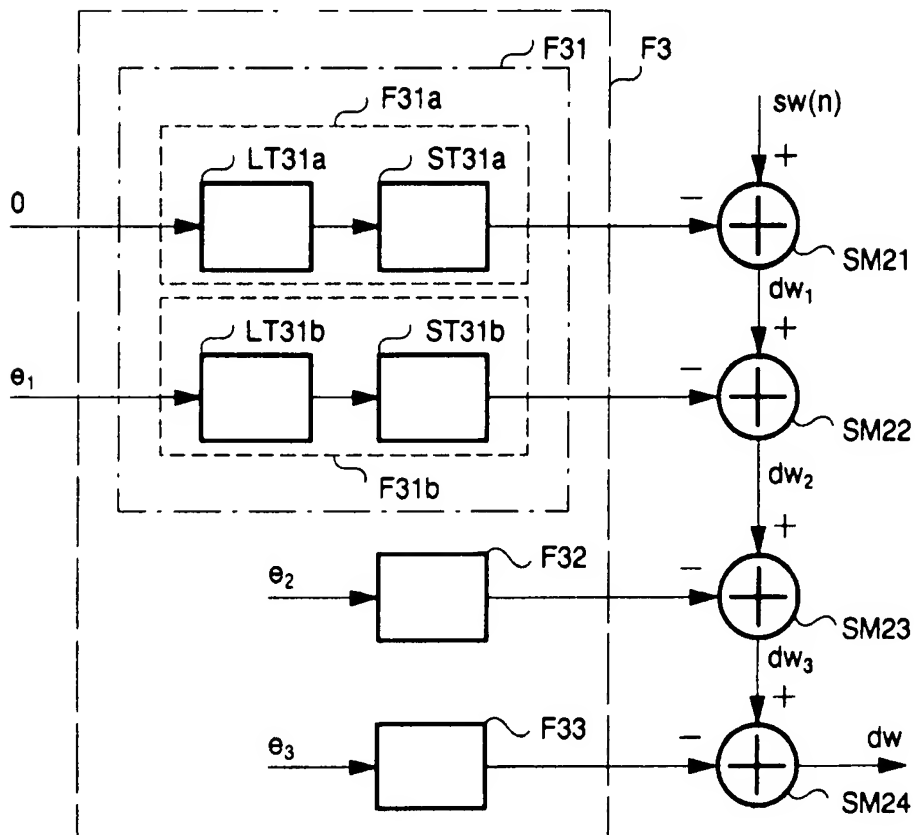


FIG. 4

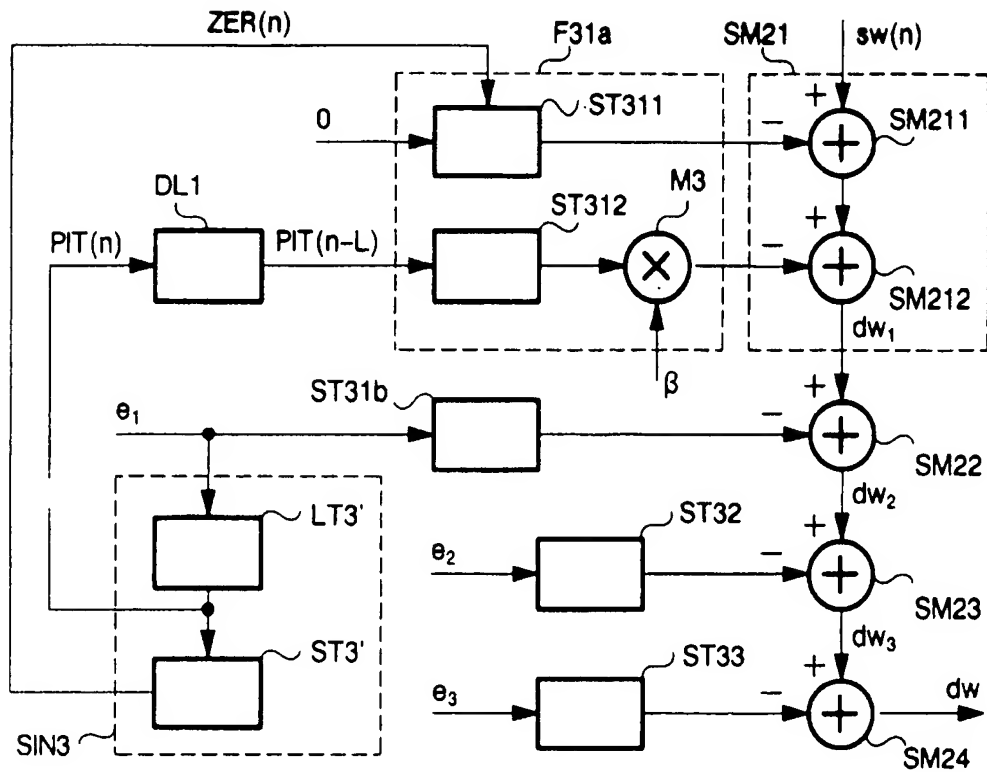


FIG. 5

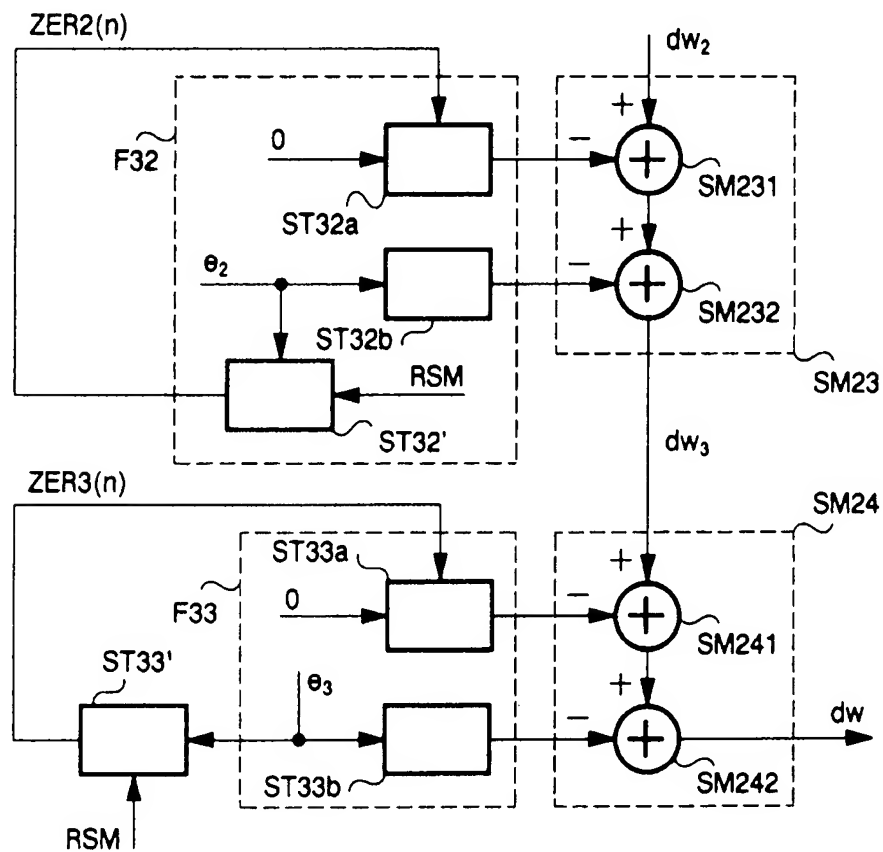


FIG. 6